Санкт-Петербургский государственный университет

Выпускная квалификационная работа

Направление «01.06.01 Математика и механика»

Обазовательная программа «МК.3001.2017 Математика»

Профессиональная траектория «Статистическое моделирование»

Скурат Евгения Петровна

Статистические многокритериальные методы анализа данных с приложением в фармакологии и генетике

Научный руководитель

к. ф.-м. н., доцент Н. П. Алексеева

Рецензент

к. ф.-м. н. П. В. Ананьевская

Санкт-Петербург

2021

Saint Petersburg State University

Final qualifying work

Direction «01.06.01 Mathematics and Mechanics»

Educational program «MK.3001.2017 Mathematics»

Professional trajectory «Statistical modeling»

Skurat Evgeniia Petrovna

Statistical multicriteria methods of analysis data with application in pharmacology and genetics

Scientific Supervisor

Candidate of Physico-Mathematical Sciences,

Associate Professor N. P. Alexeeva

Reviewer

Candidate of Physico-Mathematical Sciences

P. V. Ananievskaia

Saint Petersburg

2021

Оглавление

Введение

Статистическая задача сравнения одной зависимой переменной с набором нескольких независимых дихотомических переменных является актуальной, особенно, когда влияние различных факторов на зависимую переменную изучается отдельно и все возможные взаимосвязи незначительны. Становится очевидным, что отдельных факторов зачастую недостаточно для описания группы риска. В данной ситуации, в случае учитывая множества факторов, возникает проблема уменьшения размерности, которая означает поиск нескольких функций факторов с наименьшей потерей информации. Модели таких функций могут быть разными. В рамках данной работы мы рассматриваем и применяем модели симптом-синдромные (Алексеева Н.П. 2013). Для данных моделей предикат выражается в виде линейных комбинаций над полем $\mathbb{F}_2$, которые образуют конечное проективное пространство. Если построить конечное проективное пространство для $2^k - 1$ различных невырожденных умножений без повторений, то получим полиномы Жегалкина, которые и описывают все виды логических функций - все возможные комбинации логических операций этих k переменных: сложение, умножение, отрицание.

Известно, что каждая логическая функция может быть представлена в форме полинома Жегалкина уникальным образом, поэтому, используя их для итерации, можно найти логическую функцию, которая наилучшим образом описывает группу риска.

К сожалению, существует проблема в сложности расчетов, которая приводит к вводу ограничения: будем рассматривать порядок $3 - 4$ зависимых факторов. Отметим, что этого порядка уже достаточно для определения группы риска, которая описывается логической комбинацией факторов.

Данный метод был изучен и практически применен в главе 1 для выявления генетических факторов риска у пациентов с синдромом алкогольной зависимости, получающих терапию алкогольной зависимости (Санкт-Петербургский психоневрологический научно-исследовательский институт им. В.М. Бехтерева). При анализе симптомов выживания использовался тест Э. Уилкоксона Э. А. Гехана (1975).

В рамках анализа данных, описанных в главе 1, не удалось выявить никаких значимых отличий между индексом тяжести зависимости (психиатрический статус, употреб-

ление наркотиков) и генетическими факторами у пациентов с синдромом алкогольной зависимости. В связи с этим, появилась потребность обратиться к двумерному гамма-распределению для проверки того, а нет ли значимых отличий в динамике в разных группах, данному подходу посвящена глава 2.

Следующим шагом возникла идея объединить оба подхода к исследованию данных и совместить их, что было выполнено во второй главе в разделе 2.8: сначала посредством перебора всех возможных симптомов и суперсимптомов найти тот, которые согласно модели двумерно гамма-распределения даст значимые отличия в разных группах в динамике (в нашем случае динамике по времени выбывания из программы).

Завершающая часть работы (глава 3) посвящена еще одной популярной проблеме - анализ неполных данных. В работе рассмотрена идея анализа неполных данных без удаления или замены пропусков. Идея предлагаемого метода заключается в том, чтобы вместо одной дискриминантной функции, построенной сразу по всем независимым переменным, рассмотреть совокупность наиболее значимых частичных дискриминантных функций. Откуда возникла задача выражения полной дискриминантной функции через частные.

Глава 1

Симптомный анализ

1.1. Симптом и супер-симптом

Рассмотрим некоторый случаный вектор $X = (X_1, \ldots, X_m)^{\mathrm{T}}$ с компонентами, принимающими значения 0 или 1. Обычно 0 и 1 означают отсутствие и наличие факторов соответственно. Новая переменная $X_i + X_j \pmod 2$ означает наличие любого из двух факторов при отсутствие другого.

Определение 1.1.1. *Пусть* $\tau = (t_1, \ldots, t_k) \subseteq (1, 2, \ldots, m)$. *Тогда* $X_\tau = \sum_{i=1}^{k} X_{t_i} \pmod 2$ *называется симптомом* X_τ *ранга* $k > 2$.

Например, $\tau = (1, 3, 4)$, где $X_\tau = X_1 + X_3 + X_4 \pmod 2$. Компоненты вектора X являются тривиальными симптомами ранга 1. Определим симптом нулевого ранга вырожденным,он принимает значение 0 с вероятностью 1.

Определение 1.1.2. *Пусть определены переменные* $X_1, \ldots, X_m$, *и для* $k \leq m$ *подпространство* $\tau = (t_1, \ldots, t_k) \subseteq (1, 2, \ldots, m)$.

Обозначим результат умножения как $X^\tau = X_{t_1} \cdot \ldots \cdot X_{t_k}$. *Если* $\{\tau_j\}_{j=1}^{L}$, *где* $\tau_j \subseteq (1, 2, \ldots, m)$, $\tau_i \neq \tau_j$ *для* $i \neq j$, *тогда* $\sum_{i=1}^{L} X^{\tau_i} \pmod 2$
называется супер-симптомом.

Например, $X_1 + X_2 + X_1X_2 \pmod 2$ и $X_1X_2 + X_1X_3 + X_2X_3 \pmod 2$ являются супер-симптомами. Первый означает наличие X_1 или X_2, или обоих вместе, второй - налчие не менее двух из трех заданных факторов. В качестве альтернативы симптомы и супер-симптомы могут быть определены с помощью специальной параметризации дихотомических векторов $(\alpha_1, \ldots, \alpha_m)$ and $(\beta_1, \ldots, \beta_M)$, где $M = 2^m - 1$, $\alpha_k, \beta_i \in \{0, 1\}$, все компоненты вектора не равны нулю одновременно: $\sum_{k=1}^{m} \alpha_k \neq 0$, $\sum_{i=1}^{M} \beta_i \neq 0$. Введем специальную параметризацию $a = \sum_{k=1}^{m} \alpha_k 2^{k-1}$ и $b = \sum_{i=1}^{M} \beta_i 2^{i-1}$. Далее симптомы и супер-симптомы определяются для $a = 1, \ldots, M = 2^m - 1$ and $b = 1, 2, \ldots, 2^M - 1$ соответственно

в виде линейных комбинаций и полиномов виде

$$G_a(X_1,\ldots,X_m)=\sum_{k=1}^{m}\alpha_k X_k \pmod 2,$$
$$F_b(X_1,\ldots,X_m)=\sum_{a=1}^{M}\beta_a\prod_{k=1}^{m}X_k^{\alpha_k} \pmod 2. \tag{1.1}$$

1.2. Полиномы Жегалкина

Многочлен Жегалкина определяется как многочлен с коэффициентами вида 0 и 1, где операция сложения и умножения выполняются над полем $\mathbb{F}_{\not 2}$:

$$f(X_1,\ldots,X_n)=\sum_{\tau\subseteq(1,\ldots,n)}a_\tau\prod_{i\in\tau}X_i \pmod 2,\quad a_\tau\in 0,1\,. \tag{1.2}$$

В качестве алгоритма полного поиска всех возможных многочленов Жегалкина предлагается рекурсивно использовать импульсную последовательность с несколькими операциями $(\cdot)$ в виде

$$P_1(X_1|\cdot)=(X_1),\quad P_2(X_1,X_2|\cdot)=(X_1,X_2,X_1\cdot X_2),\ldots$$
$$P_i(X_1,\ldots,X_i|\cdot)=(P_{i-1},X_i,P_{i-1}\cdot X_i),\ X_i\notin P_{i-1},\ i>2.$$

Суперпозиция сначала умножения $(*)$, а зачем сложения $(\oplus)$ над полем $\mathbb{F}_2$ соответственно, обеспечивает простой способ построения всех видов многочленов Жегалкина как $P_m(P_n(X_1,\ldots,X_n|*)|\oplus)$, где $m=2^n-1$.

Цель выбора функции - выбрать подмножество переменных, игнорируя функции с менее важной информацией.

1.3. Алгоритм отбора

Пусть дихотомические переменные $X_1,\ldots,X_m$ включены в некоторый статистический тест. Рассмотрим все возможные комбинации X_{t_1},X_{t_2},X_{t_3}, зачем вычислим F_b для $m=3$ и энтропии $H(F_b)$ соответствующих распределений. Обозначим p-value теста через p. Любой фактор считается значимым при $p<0.05$ с поправкой на множественные сравнения и $H(F_b)>0.05$. Таким образом, мы выбираем наиболее значимый суперсимптом из всех трех возможных переменных. Аналогичная процедура для четырех переменных занимает намного больше времени. Однако частный случай $m=3$ позволяет сделать интересные выводы.

Каждый из $127 = 2^7 - 1$ возможных супер-симптомов, построенных из трех переменных a, b, c, может быть выражен как полиномиальный модуль 2 и как комбинация логических операций. Почти половина супер-симптомов (63 из 127) имеют довольно простую интерпретацию, это может быть 23 вида умножения $\alpha^{k_1}\beta^{k_2}\gamma^{k_3}$, где $\alpha \in \{a, \bar{a}\}, \beta \in \{b, \bar{b}\}, \gamma \in \{c, \bar{c}\}$. Степени k_1, k_2, k_3 принимают значения 0 или 1. Число таких выражений равно $23 = \sum_{j=1}^{3} C_3^j 2^j - 3$ потому что a и $\bar{a}$ означают один некоторый симптом в точности до кодирования перестановки и при $j = 1$ мы имеем только три выражения a, b, c а не шесть (a, b, c, $\bar{a}$, $\bar{b}$, $\bar{c}$).

кроме того, у нас есть многочлены вида $\alpha\beta + \alpha\gamma + \beta\gamma \pmod 2$, что означает наличие двух или более факторов. Таких многочленов четыре, а не восемь, потому что если заменить две переменные, например, a и b на противоположные, тогда полученное выражение отличается на 1 от многочлена, который получается, когда оставшаяся переменная c заменяется на противоположную, $\bar{a}\bar{b} + \bar{a}c + \bar{b}c \pmod 2 = ab + a\bar{c} + b\bar{c} + 1 \pmod 2$. И если мы заменяем все переменные на противоположные в многочлене, затем получаем противоположный многочлен, $\bar{a}\bar{b} + \bar{a}\bar{c} + \bar{b}\bar{c} \pmod 2 = ab + ac + bc + 1 \pmod 2$.

Мы можем построить $24 = C_3^1 \cdot 2^3$ супер-симптомы вида $\alpha(\beta + \gamma + \beta\gamma)$, соответствующие наличию факторов α одновременно с β или γ. Наконец, мы можем добавить $12 = C_3^1 \cdot 2^2$ выражений вида $\alpha\beta + (\alpha + 1)\gamma$, соответствующих наличию фактора α вместе с β или противоположному $\bar{\alpha}$ с фактором γ. Получаем так же четыре выражения вемсто восьми, потому что для всех $\alpha \in \{a, \bar{a}\}$ имеем $\alpha b + \bar{\alpha}c = \alpha\bar{b} + \bar{\alpha}\bar{c} + 1$ и $\alpha\bar{b} + \bar{\alpha}c = \alpha b + \bar{\alpha}\bar{c} + 1$.

Иногда несколько супер-симптомов имеют сопоставимые p-values. В этом случае становится возможным выбрать в качестве именительного представителя фактор, который более доступен для интерпретации.

1.4. Симптомный анализ прикладных данных из генетики

В статистическом анализе отбор набора данных или конкретной переменной является активно развивающейся областью исследований. Прежде чем приступать к применению метода классификации данных, можно попытаться применить методы отбора переменных, чтобы минимизировать количество признаков в наборе данных. Поэтому цель, которая ставиться в данной части - выполнить отбор подмножества переменных, игнорируя функции с менее важной информативностью.

1.4.1. Описание эксперимента

Исследование проводилось на базе отделения наркологии Национального медицинского научно-исследовательского центра психиатрии и неврологии им. В. М. Бехтерева в период 2013 – 2017. В рамках двойного слепого плацебо-контролируемого исследования 100 пациентов с синдромом алкогольной зависимости (МКБ-10) были случайным образом распределены (рандомизированы) на 2 группы:

1. пациенты основной группы (50 человек) получали прегабалин в дозе 150 мг / день (ночью)

2. пациенты группы сравнения (контрольная группа) (50 человек) получали плацебо идентичного вида.

Препарат для исследования назначался на 3 месяца (12 недель), в течение которых испытуемые должны были посещать исследовательский центр еженедельно для контроля ремиссии, соблюдения режима приема лекарств (при наличии флуоресцентного маркера рибофлавина в моче), а также для психометрические оценки. Кроме того, у пациентов были взяты образцы крови.
По техническим причинам 86 пациентов были доступны для анализа, образцы крови оставшихся пациентов были потеряны или выделение ДНК и генотипирование было невозможно. Пациенты были исключены из исследования в случае рецидива алкоголизма, который рассматривался как возобновление массового ежедневного (пьянства) употребления алкоголя в течение четырех или более дней подряд (в соответствии с международными стандартами «тяжелого алкоголя» - употребление алкоголя 5 и более стандартных порций алкоголя в день для мужчин и 4 или более для женщин), а также в случае пропуска трех или более посещений подряд. Различия в продолжительности ремиссии в программе лечения у носителей разных полиморфных вариантов генов и их комбинаций были выполнены с использованием анализа выживаемости Каплана-Мейера. Значимость различий в кривых выживаемости оценивали с использованием критерия Уилкоксона Гехана. Взаимосвязь исходов и отдельных полиморфных вариантов генов оценивали независимо от группы терапии.

В качестве шкалы времени рассматривались - время до выхода из программы (время до выхода из программы по любой причине, рецидив, нарушение условий участия).

Целью исследования является разработка математического метода выявления совокупности генетических факторов, значимо влияющих на тяжесть алкогольной зависимости зависимости. Методы решения поставленной задачи: построение полиномов Жегалкина над конечным полем факторов рецессивности генов (симптомов) и использование их в качестве дополнительного с терапией фактора в анализе данных типа времени жизни с разными формами цензурирования,

Дихотомическая переменная со следующими значениями считалась шкалой исключения:

- 1 (выход из программы лечения),

- 0 (завершение программы лечения).

Для анализа была использована рецессивная модель - три генотипа для каждого полиморфного локуса были объединены в две группы:

- 0 (носители одного аллеля в гомозиготном состоянии),

- 1 (все остальные генотипы).

Рассмотрим состав генетической панели, выбранной в данном исследовании, которая включает в себя код, рецессивную модель и ее расшифровку:

1. $G_1, [CC, (CT, TT)]$ - SNP в гене $DRD2$ гена рецептора допамина $D2$ $rs1799732$ $DRD2$ (ген рецептора допамина типа 2) обнаружен в значительных количествах в лимбической системе мозга и играет важную роль в функционировании центральной нервной системы. Нейрофизиологические показатели двигательного и нейрокогнитивного дефицита у пациентов с психотическими расстройствами имеют различные ассоциации с генами, регулирующими дофаминовую и глутаматную системы. Полиморфизм в гене $DRD2$ связан с пониженной концентрацией или с тяжелым алкоголизмом.

2. $G_2, [TT, (CT, CC)]$ - SNP в гене $DRD4$ гена рецептора допамина $D4$ $rs1800955$ DRD4 является основным акцептором нейронного импульса в системе нейротрансмиттеров допамина, он расположен на конце нейрона, который получает нервный импульс, и опосредует эффекты дофамина в качестве нейротрансмиттера. $DRD4$

экспрессируется на высоких уровнях в префронтальной коре и является доминирующим рецептором DA, локализованным в этой области мозга. Носитель минорного аллеля C связан с менее эффективным лечением серотонина и повышенной восприимчивостью к поиску новинок.

3. $G_3, [AA, (AG, GG)]$ - SNP в гене опиоидных μ-рецепторов $OPRM1$ $rs1799971$ Опиоидные рецепторы μ-типов ($OPRM1$) являются наиболее важным эффектором эффекта усиления опиоидов. Аллель $rs1799971$ (G) в экзоне 1 гена μ-опиоидного рецептора OPRM1 вызывает замену нормальной аминокислоты в остатке 40, аспарагина (Asn), на аспарагиновую кислоту (Asp). Носители по крайней мере одного аллеля $rs1799971$ (G), по-видимому, имеют более сильную тягу к алкоголю, чем носители двух аллелей $rs1799971$ (A)

4. $G_4, [TT, (CT, CC)]$ - SNP в гене рецепторов GABA- $rs567926$ Альфа-рецепторы гамма-аминомасляной кислоты GABA-α представляют собой ионотопный рецептор и лиганд-активированный ионный канал. Эндогенный лиганд ГАМК, при связывании которого происходит гиперполяризация мембраны нейрона, что является основой ингибирующего эффекта ГАМК. Имеются данные о связи ряда полиморфизмов генов этих субъединиц с алкогольной зависимостью.

1.4.2. Описание результатов

Симптомный, супер-симптомный метод позволил определить, что пациенты лучше всего различаются по продолжительности ремиссия с помощью сложных генетических факторов. Во-первых, выделяется группа пациентов с генотипами $[G_1(CC), G_3(AG, AA), G_4(TT)]$ or $[G_1(CC), G_3(GG), G_4(CC, CT)]$ который может быть описан супер-симптомом вида

$$S = (1 + G_1)(G_3 + G_4) \pmod 2 .$$

Это составной фактор S ($S = 1$) означает налчие одного из двух факторов риска G_3, G_4 и отсутствие третьего G_1. Пациенты в этой группе ($S = 1$) больше всего отличаются от остальных и имеют наименьшее количество рецидивов. Среднее время, проведенное в программе, равно 9.92 для $S = 1$ и 6.56 для $S = 0$, значение критерия Гехана-Уилкоксона равно $p = 1.8 \cdot 10^{-5}$. Вероятность выхода равна $p_1 = 0.35(37)$ при $S = 1$ по сравнению

с $p_0 = 0.82(45)$ в группе с $S = 0$, значение точного критерия Фишера равно $p = 2.8 \cdot 10^{-5}$.

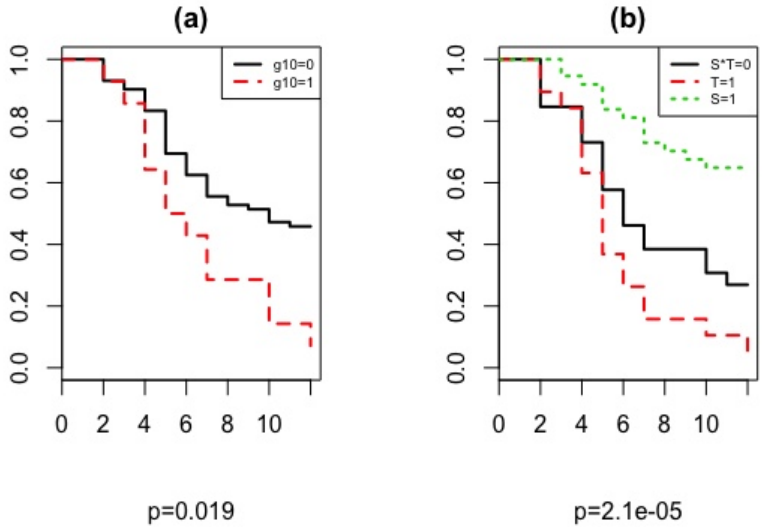

Рис. 1.1. Кривые выживания Каплана-Мейера для продолжительности безрецидивного периода с факторами: (a) G_1, (b) комбинация $S = (1 + G_1)(G_3 + G_4) \pmod 2$ и $T = G_1G_3 + G_1G_4 + G_3G_4 \pmod 2$

В сочетании с генами фактора S могут быть идентифицированы другие факторы, например,

$$T = G_1G_3 + G_1G_4 + G_3G_4 \pmod 2.$$

Этот составной фактор T ($T = 1$) означает наличие как минимум двух из трех факторов риска. Среднее время ремиссии равно $5,68$ при $T = 1$ и $8,8$ при $T = 0$, значение критерия Гехана Уилкоксона равно $p = 0,0003$. Это не лучший вариант, но описать группу с менее эффективным лечением мы можем. Результаты представлены на конференции [6] и опубликованы в [7].

Глава 2

Модель двумерного гамма-распределения

К модели двумерного гамма-распределения пришлось обратиться в связи с тем, что по анализу усредненных психологических характеристик не было выявлено никаких значимых отличий в генетической палитре больных с алкогольной зависимостью.

2.1. Обоснование модели

Рассмотрим $X_i \sim G(\alpha_i, \beta_i, \gamma_i), \quad i = 1, \ldots, k$ независимые гамма-распределенные случайные величины.

Определение 2.1.1. *Пусть определены Z_i такие, что $Z_i = \frac{\beta_i}{\beta_0} X_0 + X_i, i = 1, \ldots, k$. Тогда совместное распределение вектора $Z = (Z_1, \ldots, Z_k)$ называется многомерным гамма распределением с плотностью распределения*

$$f(z_1, \ldots, z_k) = \frac{(z_1 - \gamma_1)^{\alpha_1 - 1} \cdot (z_2 - \gamma_2)^{\alpha_2 - 1} \ldots (z_k - \gamma_k)^{\alpha_k - 1} \cdot e^{-((z_1 + \ldots + z_k) - (\gamma_1 + \ldots + \gamma_k))/(\beta_1 \cdot \ldots \cdot \beta_k)}}{\prod_{i=1}^{k} \beta_i^{\alpha_i} \Gamma(\alpha_i)},$$

где $\alpha_i, \beta_i, \gamma_i$ параметры формы, масштаба и местоположения соответственно, для которых $\alpha_i > 0,\ \beta_i > 0,\ \gamma_i,\ \gamma_i < z_i, i = 2, \ldots, k,\ z_k < \infty$, и ноль иначе.

Теорема 2.1.1. *Определим Z_i такие, что $Z_1 = X_1,\ Z_2 = X_1 + X_2, \ldots,\ Z_k = X_1 + \ldots + X_k$, тогда совместное распределение вектора $Z = (Z_1, \ldots, Z_k)$ является многомерным гамма распределением с плотностью распределения*

$$\begin{aligned} f(z_1, \ldots, z_k) = {} & \frac{(z_1 - \gamma_1)^{\alpha_1 - 1}}{\beta^{\alpha_k^*} \prod_{i=1}^{k} \Gamma(\alpha_i)} \\ & \cdot (z_2 - z_1 - \gamma_2)^{\alpha_2 - 1} \ldots (z_k - z_{k-1} - \gamma_k)^{\alpha_k - 1} \\ & \cdot e^{-(z_k - (\gamma_1 + \ldots + \gamma_k))/\beta}, \end{aligned} \tag{2.1}$$

где $\alpha_i > 0,\ \beta > 0,\ \gamma_i,\ z_{i-1} + \gamma_i < Z_i, i = 2, \ldots, k,\ z_k < \infty,\ \gamma_1 < z_1,\ \alpha_k^* = \alpha_1 + \ldots + \alpha_k$, и ноль иначе.

Определим теперь многомерную гамма-распределенную случайную величину в терминах характеристической функции.

$$\psi_z(u_1, \ldots, u_n) = \prod_{i=1}^{n} \frac{\psi_{z_j}(u_j + \sum_{k=j+1}^{n} \beta_{j_k} u_k)}{\psi_{z_j}(\sum_{k=j+1}^{n} \beta_{j_k} u_k)}, \tag{2.2}$$

где $\psi_{z_j}(u_j) = (1 - iu_j/a_j)^{-e_j},\ j = 1, \ldots, n,\ i = \sqrt{-1},\ \beta_{jk} \geq 0,\ \ a_j \geq \beta_{jk}a_k > 0,\ j < k = 1, \ldots, n,\ 0 < e_1 \leq e_2 \leq \ldots \leq e_n$.

С помощью этого определения изучаются различные свойства, но явный вид плотности оценивается только для двумерного случая.

Дополнительно рассмотрим производящую функцию моментов для величины $Z = (Z_1, \ldots, Z_k)$ и параметров $|t_i + t_{i+1} + \ldots + t_k| < 1/\beta i = 1, \ldots, k$

$$M_Z(t) = E(e^{t_1 Z_1 + \ldots + t_k Z_K}) =$$

$$= \frac{e^{\gamma_1(t_1+\ldots+t_k)}}{[1-\beta(t_1+\ldots+t_k)]^{\alpha_1}} \frac{e^{\gamma_2(t_2+\ldots+t_k)}}{[1-\beta(t_2+\ldots+t_k)]^{\alpha_2}} \cdots \frac{e^{\gamma_k t_k}}{[1-\beta t_k]^{\alpha_k}}. \tag{2.3}$$

Теперь на основании введенных определений нетрудно получить следующие свойства:

1. Предельное распределение величины Z_i является гамма $Z_i \sim G(\alpha_i^*, \beta, \gamma_i^*),\ i = 1, \ldots, k,\ \ \alpha_i^* = \alpha_1 + \ldots + \alpha_i, \gamma_i^* = \gamma_1 + \ldots + \gamma_i$.

2. среднее и дисперсия определены соответственно $E(Z_i) = \beta\alpha_i^* + \gamma_i^*,\ \ D(Z_i) = \beta^2\alpha_i^*$

Если упросить рассматриваемую модель, положив $Z_i \sim G(\alpha_0 + \alpha_i, \beta_i)$ и принять все параметры местоположения γ_i равными нулю, то исходя из [5] получаем моменты случайной величины Z_i и смешанные моменты соответственно:

$$E(Z_i^m) = \sum_{r=0}^{m} C_m^r \alpha_0(\alpha_0+1)\ldots(\alpha_0+r-1)\alpha_i(\alpha_i+1)\ldots(\alpha_i+m-r-1)\beta_i^m \tag{2.4}$$

$$E(Z_i^m Z_j^n) = \sum_{r=0}^{m}\sum_{s=0}^{n} C_m^r C_n^s \left(\frac{\beta_i}{\beta_0}\right)^r \left(\frac{\beta_j}{\beta_0}\right)^s M_0^{(r+s)} M_i^{(m-r)} M_j^{(n-s)}, \tag{2.5}$$

где $M_i^{(m)} = \alpha_i(\alpha_i+1)\ldots(\alpha_i+m-1)\beta_i^m$

2.2. Случай двумерного гамма-распределения

Рассмотрим теперь более частный случай модели - двумерное распределение с параметрами масштаба равными единице.

Пусть даны три независимые гамма-распределенные случайные величины X_0, X_1, X_2 с параметром масштаба, равным 1, и параметрами экстенсивности, равными $\lambda_0, \lambda_1, \lambda_2$. Построим из них две независимые случайные величины $Y_1 = X_0 + X_1$ и

$Y_2 = X_0 + X_2$. В силу аддитивности гамма-распределения полученные случайные величины будут гамма-распределенными с параметрами формы $\lambda_0 + \lambda_1$ и $\lambda_0 + \lambda_2$. Из ранее представленных моментов случайных гамма-распределенных величин получаем

$$E(Y_1, Y_2) = E(X_0 + X_1)(X_0 + X_2) = \lambda_0(\lambda_0 + \lambda_1) + \lambda_0(\lambda_1 + \lambda_2) + \lambda_1\lambda_2 \tag{2.6}$$

$$cov(Y_1, Y_2) = E(Y_1Y_2) - EY - 1EY_2 =$$
$$= \lambda_0(\lambda_0 + \lambda_1 + \lambda_2 + 1) + \lambda_1\lambda_2 - (\lambda_0 + \lambda_1)(\lambda_0 + \lambda_2) = \lambda_0 \tag{2.7}$$

Откуда следует, что величины являются коррелированными с коэффициентом корреляции

$$\rho = \rho(Y_1, Y_2) = \frac{\lambda_0}{\sqrt{(\lambda_0 + \lambda_1)(\lambda_0 + \lambda_2)}}. \tag{2.8}$$

Теорема 2.2.1. *Если рассмотреть теперь две коррелирующие гамма-распределенные случайные величины $Y_1,\ Y_2$ с коэффициентом корреляции равным ρ, параметрами масштаба, равными 1, и с параметрами формы равными $\Lambda_1 = \lambda_0 + \lambda_1$ и $\Lambda_2 = \lambda_0 + \lambda_2$, то тогда параметры скрытой экстенсивности имеют вид:*

$$\lambda_0 = \rho\sqrt{\Lambda_1\Lambda_2}, \tag{2.9}$$
$$\lambda_1 = \Lambda_1 - \rho\sqrt{\Lambda_1\Lambda_2}, \tag{2.10}$$
$$\lambda_0 = \Lambda_2 - \rho\sqrt{\Lambda_1\Lambda_2},. \tag{2.11}$$

2.3. Плотность двумерного гамма-распределения

Для возможности расчета оценок параметров $\lambda_0,\ \lambda_1,\ \lambda_2$ по методу максимального правдоподобия необходимо получить формулу плотности двумерного гамма-распределения. Рассмотрим совместную плотность независимых гамма-распределенных случайных величин $X_0,\ X_1,\ X_2$ вида:

$$f(x_0, x_1, x_2) = C_1 x_0^{\lambda_0 - 1} x_1^{\lambda_1 - 1} x_2^{\lambda_2 - 1} e^{-(x_0 + x_1 + x_2)}, \tag{2.12}$$
$$C_1 = \frac{1}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)}, \quad x_0, x_1, x_2 > 0$$

Определим совместную плотность величин Y_1 и Y_2. Для этого введем новые переменные

$$\begin{cases} u = x_0 + x_1, \\ v = x_0 + x_2, \\ t = x_0, \end{cases} \tag{2.13}$$

Тогда получим

$$\begin{cases} x_1 = u - t, \\ x_2 = v - t, \\ x_0 = t, \end{cases} \tag{2.14}$$

Таким образом, подставляя новые переменные и интегрируя по переменной t, получаем формулу плотности:

$$f(u,v) = \frac{1}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)} \int_0^{min(u,v)} t^{\lambda_0-1}(u-t)^{\lambda_1-1}(v-t)^{\lambda_2-1}e^{-(u+v-t)}dt \tag{2.15}$$

Если положить $u < v$, разложить экспоненту в ряд Тейлора $e^t = \sum_{n=0}^{\infty} \frac{t^n}{n!}$ и вынести из под знака интеграла все независящие от t значения, тогда, применив свойства интеграла, получим:

$$f(u,v) = \frac{1}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)} e^{-(u+v)} \sum_{n=0}^{\infty} \frac{1}{n!} \int_0^u t^{u+\lambda_0-1}(u-t)^{\lambda_1-1}(v-t)^{\lambda_2-1}dt. \tag{2.16}$$

Сделаем еще одну замену $t = su$, тогда

$$f(u,v) = \sum_{n=0}^{\infty} \frac{u^{n+\lambda_0+\lambda_1-1}v^{\lambda_2-1}}{n!} \int_0^1 s^{n+\lambda_0-1}(1-s)^{\lambda_1-1}(1-\frac{su}{v})^{\lambda_2-1}ds, \tag{2.17}$$

где $C = \frac{e^{-(u+v)}}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)}$.

Для случая $v < u$ вывод формулы идентичен с точностью до перестановки соответствующих показателей степеней. На основании выполненных преобразований приходим к гипергеометрической функции Гаусса и выражению для плотности двумерного гамма-распределения:

$$f(u,v) = \begin{cases} C\sum_{n=0}^{\infty} \frac{u^{n+\lambda_0+\lambda_1-1}v^{\lambda_2-1}}{n!} \frac{\Gamma(n+\lambda_0)\Gamma(\lambda_1)}{\Gamma(n+\lambda_0+\lambda_1)} \cdot {}_2F_1(1-\lambda_2, n+\lambda_0, n+\lambda_0+\lambda_1, \frac{u}{v}), & u < v, \\ C\sum_{n=0}^{\infty} \frac{u^{\lambda_1-1}v^{n+\lambda_0+\lambda_2-1}}{n!} \frac{\Gamma(n+\lambda_0)\Gamma(\lambda_2)}{\Gamma(n+\lambda_0+\lambda_2)} \cdot {}_2F_1(1-\lambda_1, n+\lambda_0, n+\lambda_0+\lambda_2, \frac{v}{u}), & u > v, \end{cases}$$

где $C = \frac{e^{-(u+v)}}{\Gamma(\lambda_0)\Gamma(\lambda_1)\Gamma(\lambda_2)}$. Условие сходимости ряда ${}_2F_1(a,b,c,d)$ для обоих случаев приобретает вид $\lambda_1 + \lambda_2 > 0$

2.4. Оценка параметров по методу моментов

Используя выражение для совместной плотности и для моментов гамма-распределения, которые имеют вид: $\mu_2 = \lambda, \mu_2 = 2\lambda, \mu_4 = 3\lambda(\lambda + 2)$, центральные моменты двумерного гамма-распределения с единичным парамертом масштаба, необходимые для вычисления асимптотической дисперсии оценок параметров как функций от моментов, равны

$$\mu_{11} = \lambda_0, \tag{2.18}$$

$$\mu_{12} = \mu_{21} = 2\lambda_0, \tag{2.19}$$

$$\mu_{22} = 2\lambda_0^2 + 6\lambda_0 + (\lambda_0 + \lambda_1)(\lambda_0 + \lambda_2). \tag{2.20}$$

Пусть задана выборка наблюдений $(x_{11}, x_{21}), \ldots, (x_{1n}, x_{2n})$ с плотностью распределения 2.16. Оценим параметры двумерного гамма-распределения по методу моментов. Если считать $\overline{y_1}$, $\overline{y_2}$ выборочными средними, а m_{11} - вторым выборочным смешанным центральным моментом, то оценки по методу моментов представляются $\widehat{\lambda_0} = m_{11}$, $\widehat{\lambda_1} = \overline{y_1} - m_{11}$ $\widehat{\lambda_2} = \overline{y_2} - m_{11}$ [1]

Тогда, учитывая что $\mu_2(m_{11}) = \frac{\mu_{22} - \mu_{11}^2}{n} + o(\frac{1}{n})$, получим:

$$\mathsf{D}(\hat{\lambda_0}) = \frac{\lambda_0^2 + 6\lambda_0 + (\lambda_0 + \lambda_1)(\lambda_0 + \lambda_2)}{n} + o(\frac{1}{n}) \tag{2.21}$$

Используя 2.21 и тот факт, что $\mu_2(\bar{x_i}) = \frac{\lambda_0 + \lambda_i}{n}$, можем вычислить $\mu_{11}(\bar{x_i}, m_{11}) = \frac{2\lambda_0}{n} + o(\frac{1}{n}$, откуда получаем выражение для дисперсии оценок:

$$\mathsf{D}(\hat{\lambda_i}) = \frac{1}{n}\left((\lambda_0 + \lambda_i)(2\lambda_0 + \lambda_1 + \lambda_2) + \lambda_0(\lambda_0 + 2)\right) + o(\frac{1}{n}) \tag{2.22}$$

2.5. Построение доверительных интервалов

Для построения доверительного интервала необходимо вычислить информант второго рода. С учетом оценки двух параметров распределения, информант будет представлен матрицей следующего вида:

$$I = n \begin{pmatrix} \psi'(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}$$

После обращения матрицы на ее главной диагонали будут стоять дисперсии соответствующих параметров, а элементый вне главной диагонали соответствовать ковариации.

$$I^{-1} = \frac{\beta^2}{n(\psi(\alpha)\acute{\alpha} - 1)} \begin{pmatrix} \frac{\alpha}{\beta^2} & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \acute{\psi(\alpha)} \end{pmatrix},$$

В связи с тем, что оценки, полученные по методу максимального правдоподобия являются асимптотически нормальными, то потроим асимптотические доверительные интервалы:

$$P(\hat{\theta} - u_{1-\alpha/2} \cdot \frac{\sigma(\hat{\theta})}{\sqrt{n}} < \theta < \hat{\theta} + u_{1-\alpha/2} \cdot \frac{\sigma(\hat{\theta})}{\sqrt{n}}) = 1 - \alpha, \tag{2.23}$$

где $u_{1-\alpha/2}$ - квантиль стандартного нормального распределения, n - размер выборки. приняв уровень значимости за α, получаем доверительный интервал, в который случайная величина θ попадает с вероятностью $1-\alpha$. Извлекая корень из элементов $diag(I^{-1})$, получим стандартное отклонение, участвующее в построении доверительного интервала.

2.6. Применение модели на прикладных данных из генетики

2.6.1. Описание эксперимента

Для данных, описанных в разделе 1.4.1, дополнительно рассмотрим показатель оценки экстрапирамидных побочных эффектов по шкале Симпсона-Ангуса. Экстрапирамидные расстройства снижают качество жизни пациентов, их трудовую и социальную активность, приводят к когнитивным нарушениям. Они осложняют течение основного заболевания, увеличивая выраженность негативных, когнитивных и аффективных расстройств, и приводят к дополнительной социальной стигматизации пациентов. В некоторых случаях психопатологические проявления (высокая тревожность, негативная симптоматика и когнитивные расстройства), обычно трактующиеся как симптомы шизофрении, могут быть обусловлены экстрапирамидной симптоматикой при приеме нейролептиков.

В качестве независимой переменной был выбран генетической фактор носителей полиморфного аллеля $G_1, [CC, (CT, TT)]$ - SNP в гене $DRD2$ гена рецептора допамина $D2$ $rs1799732$.

Основной задачей является потребность оценки влияния полиморфизмов гена на выраженность побочных эффектов терапии по шкале Симпсона-Ангуса в двух точках измерения (1-й день исследования, 21-й день исследования).

Показатель оценки экстрапирамидных побочных эффектов по шкале Симпсона-Ангуса удовлетворяет двумерному гамма-распределению, так как является признаком, у которого наблюдается согласие с моделью гамма-распределения в обоих выбранных временных точках. Проверка выполнялась по критерию $\chi\hat{2}$ Пирсона со значением $p - value$:

Таблица 2.1. Результаты проверки критерия $\chi\hat{2}$ Пиросона

Временная точка	Статистика критерия	p-value
1-ый день	3.45	0.521
21-ой день	2.27	0.279

2.6.2. Описание результатов

В эксперименте были рассмотрены пары положительно коррелированных признаков, относящихся к разным временным точкам, для которых не отвергается гипотеза согласия с гамма-распределением,и оценены параметры формы $\lambda_0, \lambda_1, \lambda_2$ и параметров масштаба равным 1.

В таблице представлены оценки параметров по методу моментов и оценки максимального правдоподобия двумерного гамма-распределения оценки экстрапирамидных побочных эффектов по шкале Симпсона-Ангуса в двух группах генетической принадлежности в двух временных точках:

Таблица 2.2. Оценки параметров по методу моментов для шкалы Симпсона-Ангуса

Параметр	(CC)	(CT,TT)
λ_0	0,07	0,22
λ_1	0,15	0,29
λ_2	0,77	0,30

Оценки параметров позволили выявить значимое отличие наблюдений по параметру λ_2, который значимо ниже в группе носителей полиморфного аллеля, а значит в

Таблица 2.3. Оценки параметров по методу максимального правдоподобия для шкалы Симпсона-Ангуса

Параметр	(CC)	(CT,TT)
λ_0	1.695	1.116
λ_1	0.861	0.209
λ_2	0.100	0.202

данной группе концентрация побочных эффектов меньше.

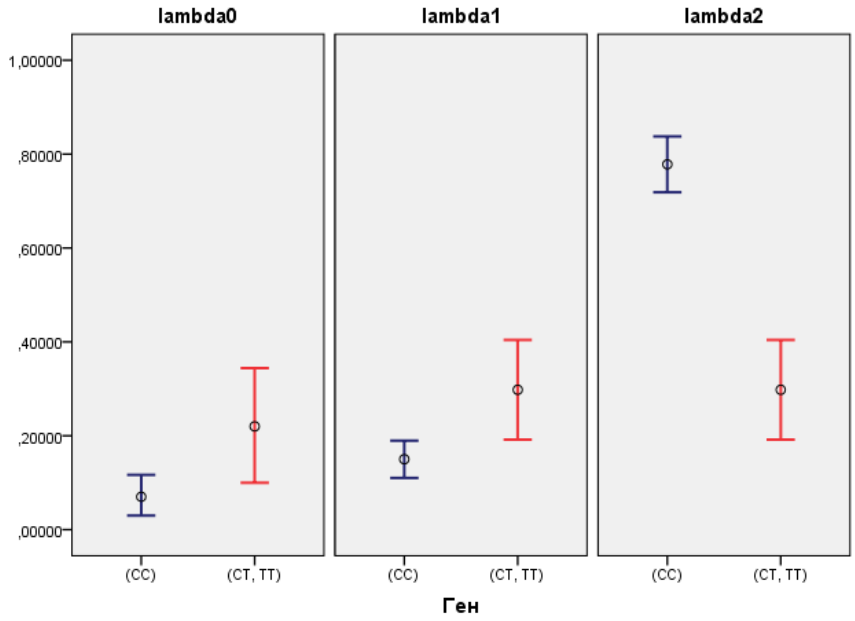

Рис. 2.1. Доверительные интервалы оценок параметров экстенсивности оценки экстрапирамидных побочных эффектов по шкале СимпсонаАнгуса в двух группах генетической палитры

2.7. Применение модели на прикладных данных из медицины

2.7.1. Описание эксперимента

Исследование проводилось на базе отделения наркологии Национального медицинского научно-исследовательского центра психиатрии и неврологии им. В. М. Бехтерева в период 2013 – 2018. В рамках двойного слепого плацебо-контролируемого исследо-

вания 306 пациентов с синдромом опиоидной зависимости были случайным образом распределены (рандомизированы) на 3 группы:

1. пациенты группы, получающие плацебо
2. пациенты группы, получающие налтрексон (блокатор опиоидных рецепторов)
3. пациенты группы, у которых налтрексон дозировался в аргонизм в виде импланта

Препарат для исследования назначался на 14 недель, в течение которых испытуемые должны были посещать исследовательский центр еженедельно для контроля ремиссии, соблюдения режима приема лекарств.
Пациенты были исключены из исследования в случае рецидива алкоголизма, который рассматривался как возобновление массового употребления опиоидов в течение четырех или более дней подряд, а также в случае пропуска трех или более посещений подряд.

В качестве шкалы времени рассматривались - время до выхода из программы по любой причине.

Целью исследования является разработка математического метода выявления значимых различий между группами пациентов в определенные моменты прохождения обследования, в случае, когда, например, посредством двухфактрного дисперсионного анализа невозможно определить в каких именно градация факторов найдены отличия. Методы решения поставленной задачи: построение модели двухфактрого гамма-распределения и анализ оценок параметров интенсивности и экстенсивности.

Рассмотрим следующие показатели состояния пациентов:

1. Аспартатаминотрансфераза (далее АСТ) - это фермент, который принимает участие в превращении аспарагиновой аминокислоты в клетке. Наибольшее количество АСТ содержится в миокарде (сердечной мышце), печени, почках и скелетных мышцах. АСТ локализуется в митохондриях и цитоплазме клеток, в связи с этим при повреждении клетки он быстро обнаруживается в крови. Быстрое нарастание концентрации аспарагиновой аминотрансферазы очень характерно для острого повреждения миокарда (например, для инфаркта). Увеличение в крови фермента наблюдается спустя 8 часов с момента поражения и достигает своего максимума спустя сутки. Снижение концентрации АСТ при инфаркте происходит на 5 день.

2. Аланинаминотрансфераза (далее АЛТ) - это внутриклеточный фермент, который участвует в метаболизме клеток, в частности – в расщеплении аминокислоты аланина. Больше всего аланиновой аминотрансферазы содержится в клетках печени, меньше – в миокарде, скелетной мускулатуре и почках. Повышение АЛТ в анализе крови происходит при любом повреждении гепатоцитов (клеток печени). Повышение фермента наблюдается уже в первые часы после повреждения и постепенно возрастает в зависимости от активности процесса и количества поврежденных клеток.

Повышение аланиновой и аспарагиновой аминотрансферазы может повышаться при многих заболеваниях (цирроз печени, алкоголизм). На практике иногда бывают случаи, когда показатели АСТ или АЛТ становятся ниже нормы. Это может случиться при тяжелом и обширном некрозе печени (например, в случае запущенного гепатита). Особенно неблагоприятным прогнозом обладает снижение уровня АСТ и АЛТ на фоне прогрессирующего нарастания билирубина. Витамин В6 необходим для синтеза АСТ и АЛТ в норме. Снижение концентрации В6 может быть связано с длительным лечением антибиотиками или голоданием. Так же на практике было доказано, что внутримышечные инъекции влияют на повышение АЛТ.

2.7.2. Описание результатов

В эксперименте были рассмотрены пары положительно коррелированных признаков, относящихся к разным временным точкам, для которых не отвергается гипотеза согласия с гамма-распределением,и оценены параметры формы $\lambda_0, \lambda_1, \lambda_2$ и параметром масштаба равным 1.

Для признака АЛТ наличие согласия с гамма распределением присутствует только в двух временных точках (на 7-й и 13-й). Именно эти точки и будем рассматривать в модели. По количеству индивидов в рассматриваемых группах получаем 2.4.

Построим теперь общую линейную модель, чтобы посмотреть наличие факта влияния взаимодейсвтия факторов времени и типа лечения на признак АЛТ. Результаты приведены в таблице 2.5 и на рисунке 2.2.

Оценим параметры экстенсивности без устранения влияния интенсивности АЛТ в группах. В таблице 2.6 и 2.7 представлены результаты оценки.

Оценим теперь параметр интенсивности 2.16 и, в случае необходимости, устраним

Таблица 2.4. Количество индивидов в группах

Временная точка	Тип лечения	Количество
7	Плацебо	25
7	Налтрексон	34
7	Налтрексон-имплант	66
13	Плацебо	9
13	Налтрексон	16
13	Налтрексон-имплант	51

Таблица 2.5. Оценки эффектов межгрупповых факторов

Источник	F	p-value
время	0,121	0,728
группа	5,62	0,004
время*группа	1,74	0,178

Таблица 2.6. Оценки параметров экстенсивности по методу МП для АЛТ

Параметр	Плацебо	Налтрексон	Налтрексон-имплант
λ_0	0,735	1,141	1,045
λ_1	1,295	0,439	0,426
λ_2	1,611	0,265	0,387

Таблица 2.7. Оценки параметров экстенсивности по методу моментов для АЛТ

Параметр	Плацебо	Налтрексон	Налтрексон-имплант
λ_0	0,4719	1,858	0,885
λ_1	3,017	0,691	0,004
λ_2	1,621	0,721	2,144

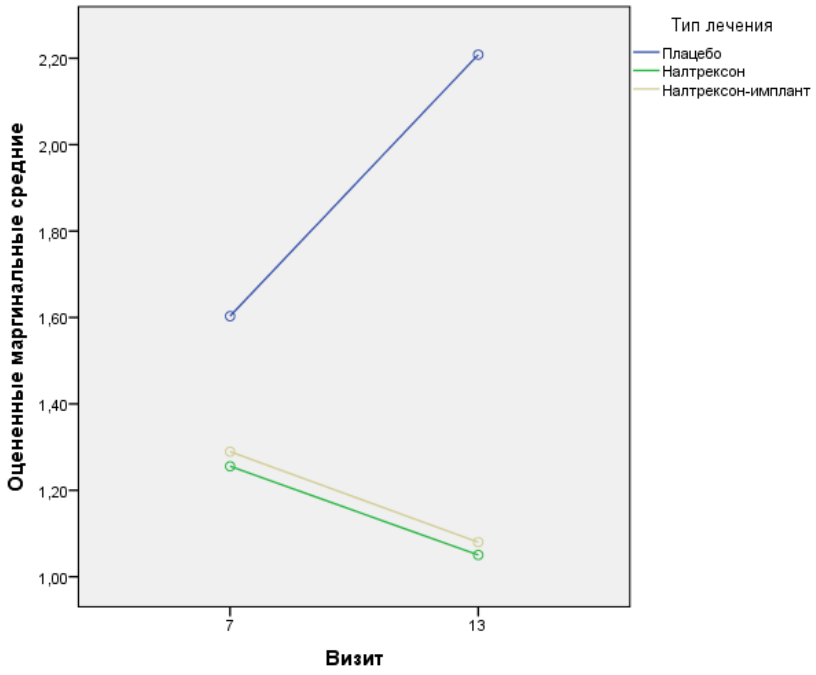

Рис. 2.2. График эффектов взаимодействия фактора времени и типа лечения

влияние интенсивности АЛТ разделив признак на полученную оценку. Деление будем выполнять в каждой временной точке и в каждой группе по типу лечения отдельно.

Таблица 2.8. Оценки параметра интенсивности АЛТ (scale)

Временная точка	Плацебо	Налтрексон	Налтрексон-имплант
7	0,447	0,544	0,615
13	0,884	0,414	0,318

По полученным результатам, видим, что значение интенсивности достаточно велико в группе Плацебо в конечной временной точке. Далее по алгоритму необходимо исключить влияние интенсивности и разделить значения признака АЛТ на соответствующие значения интенсивности.

Следующим шагом выполняем оценку параметров экстенсивности по методу моментов и методу максимального правдоподобия. В таблице 2.9 и 2.10 представлены результаты оценки.

Сравнение доверительных интервалов на уровне значимости 0,05 наблюдается на уровне группы Плацебо λ_1 и группы Налтрексон-имплант λ_2. Результаты сравнения

Таблица 2.9. Оценки параметров экстенсивности по методу МП для АЛТ

Параметр	Плацебо	Налтрексон	Налтрексон-имплант
λ_0	0,936	1,937	1,668
λ_1	2,939	0,400	0,281
λ_2	1,627	0,721	1,905

Таблица 2.10. Оценки параметров экстенсивности по методу моментов для АЛТ

Параметр	Плацебо	Налтрексон	Налтрексон-имплант
λ_0	0,4719	1,858	0,885
λ_1	3,017	0,691	0,004
λ_2	1,621	0,721	2,144

представлены на рисунке 2.3 и в таблице 2.11.

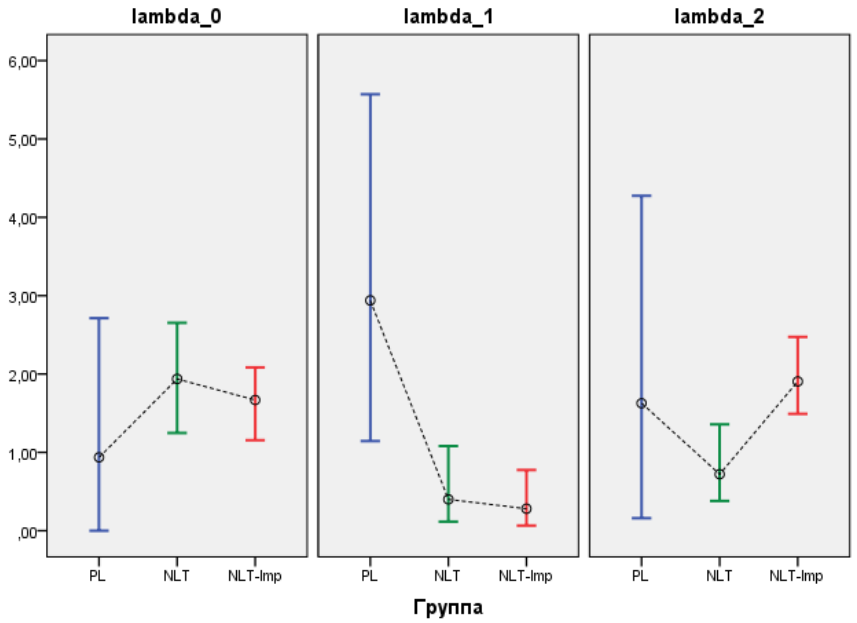

Рис. 2.3. Доверительные интервалы параметров экстенсивности АЛТ

На основании полученных результатов и увеличения λ_2 в группе налтрексон-имплант, можно сказать, что в этой группе наиболее сильно происходит появление ново-

Таблица 2.11. Доверительные интервалы параметров экстенсивности АЛТ

Группа	λ_0	λ_1	λ_2
Плацебо	(0; 2,713)	(1,144; 5,572)	(0,160; 4,274)
Налтрексон	(1,248; 2,654)	(0,117; 1,08)	(0,379; 1,357)
Налтрексон-имплант	(1,156; 2,084)	(0,065; 0,774)	(1,491; 2,473)

го АЛТ, который усиливает распад тканей печени, при этом в этой же группе маленькое значение λ_1 говорит о маленьком преобразовании показателя АЛТ с течением времени лечения.

2.8. Применение симптомного анализа и модели гамма-распределения на прикладных данных из медицины

2.8.1. Описание эксперимента

Исследование проводилось на базе отделения наркологии Национального медицинского научно-исследовательского центра психиатрии и неврологии им. В. М. Бехтерева в период 2013 – 2018. В рамках двойного слепого плацебо-контролируемого исследования 306 пациентов с синдромом опиоидной зависимости были случайным образом распределены (рандомизированы) на 3 группы:

1. пациенты группы, получающие плацебо
2. пациенты группы, получающие налтрексон (блокатор опиоидных рецепторов)
3. пациенты группы, у которых налтрексон дозировался в аргонизм в виде импланта

В качестве шкалы времени рассматривались - время до выхода из программы по любой причине. В качестве дополнительных категориальных признаков рассматривались все возможные измеренные показатели базы данных - пол, признак наличия наркотиков в крови, признак наличия алкоголя в крови так далее.

Целью исследования является разработка математического метода выявления более детальных значимых различий между группами пациентов в определенные моменты прохождения обследования, в случае, когда, например, посредством двухфактрного

дисперсионного анализа невозможно определить в каких именно градация факторов найдены отличия. Методы решения поставленной задачи: построение модели двухфактрого гамма-распределения для различных симптомов и анализ оценок параметров интенсивности и экстенсивности.

Рассмотрим следующие показатели состояния пациентов:

1. Аланинаминотрансфераза (далее АЛТ) - это внутриклеточный фермент, который участвует в метаболизме клеток, в частности – в расщеплении аминокислоты аланина.Повышение АЛТ в анализе крови происходит при любом повреждении гепатоцитов (клеток печени). У мужчин ферментный состав крови быстро реагирует на интенсивную физическую нагрузку (поднятие тяжестей, бег, спортивные тренировки), поэтому перед анализами нужно воздержаться от посещения спортзала и другой напряженной мышечной работы. У мужчин и женщин ферментный состав крови чувствителен к нервному перенапряжению, к стрессу. Так же на практике было доказано, что внутримышечные инъекции влияют на повышение АЛТ.

2. Пол - признак принадлежности пациента к мужскому или женскому полу (0 - мужчины, 1 - женщины).

2.8.2. Описание результатов

В эксперименте были рассмотрены пары положительно коррелированных признаков, относящихся к разным временным точкам, для которых не отвергается гипотеза согласия с гамма-распределением,и оценены параметры формы $\lambda_0, \lambda_1, \lambda_2$ и параметром масштаба равным 1.

Для признака АЛТ наличие согласия с гамма распределением присутствует только в двух временных точках (на 7-й и 13-й). Именно эти точки и будем рассматривать в модели. Опираясь на результаты раздела 2.6 в текущем исследовании были выбраны лишь 2 группы пациентов по типу лечения - это плацебо и налтрексон-имплант. По количеству индивидов в рассматриваемых группах получаем 2.12.

Рассмотрим симптом сочетания фактора типа терапии (только группа плацебо и налтрексон-имплант) и фактора пола. Данный симптом принимает значения:

1. 1 - для случаев плацебо у женщин и налтрекосн-имплант у мужчин, обозначим группу как PL_G,

Таблица 2.12. Количество индивидов в группах

Временная точка	Тип лечения	Пол	Количество
7	Плацебо	муж	16
7	Налтрексон-имплант	муж	44
7	Плацебо	жен	9
7	Налтрексон-имплант	жен	22
13	Плацебо	муж	9
13	Налтрексон-имплант	муж	15
13	Плацебо	жен	0
13	Налтрексон-имплант	жен	1

2. 0 - для случаев плацебо у мужчин и налтрексон-имплант у женщин, обозначим группу как $NLT - Imp\_G$.

Построим теперь общую линейную модель, чтобы посмотреть наличие факта влияния взаимодейсвтия факторов времени и построенного симптома на признак АЛТ. Результаты приведены в таблице 2.13 и на рисунке 2.2.

Таблица 2.13. Оценки эффектов межгрупповых факторов

Источник	F	p-value
время	0,33	0,321
симптом	0,512	0,0475
время*симптом	0,828	0,217

Далее по уже отработанному алгоритму необходимо исключить влияние интенсивности и разделить значения признака АЛТ на соответствующие значения интенсивности.

Следующим шагом выполняем оценку параметров экстенсивности по методу моментов и методу максимального правдоподобия. В таблице 2.14 и 2.15 представлены результаты оценки.

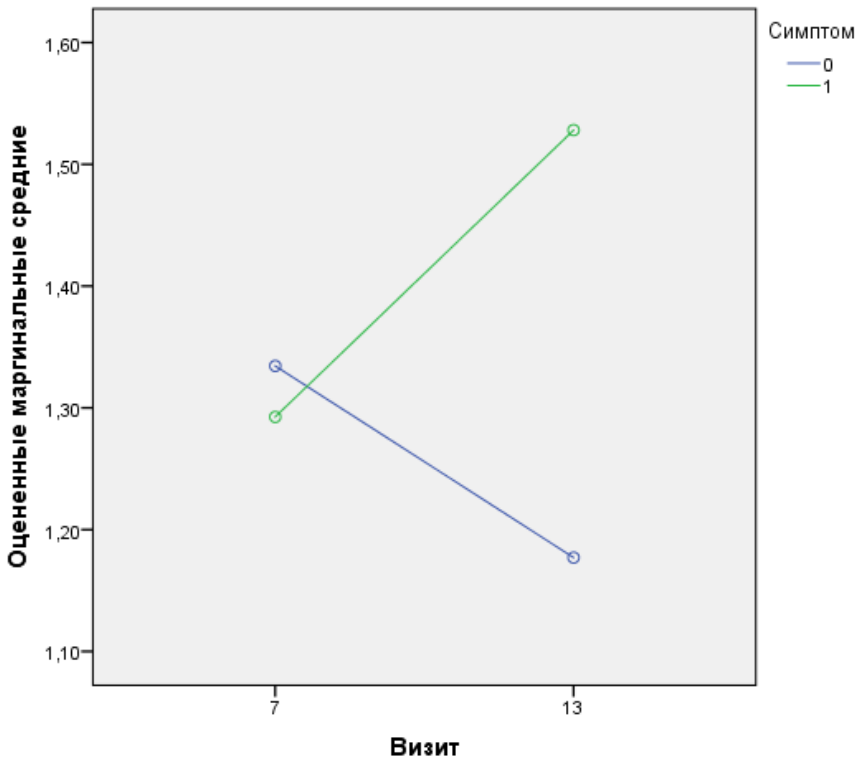

Рис. 2.4. График эффектов взаимодействия фактора времени и симптома

Сравнение доверительных интервалов на уровне значимости 0,05 не было выявлено. Результаты сравнения представлены на рисунке 2.5 и в таблице 2.16.

На основании полученных результатов и увеличения λ_2 в группе налтрексон-имплант, можно сказать, что в этой группе наиболее сильно происходит появление нового АЛТ, который усиливает распад тканей печени, при этом в этой же группе маленькое значение λ_1 говорит о маленьком преобразовании показателя АЛТ с течением времени лечения. Если обратить внимание, что в случае исследования без участния признака пола результат получался аналогичный, то можно предварительно сделать вывод, что

Таблица 2.14. Оценки параметров по методу максимального правдоподобия для АЛТ

Параметр	PL_G	$NLT - Imp\_G$
λ_0	1,071	1,776
λ_1	1,435	0,183
λ_2	1,003	2,179

Таблица 2.15. Оценки параметров по методу моментов для АЛТ

Параметр	PL_G	$NLT - Imp\_G$
λ_0	0,849	1,072
λ_1	1,578	0,109
λ_2	0,742	2,183

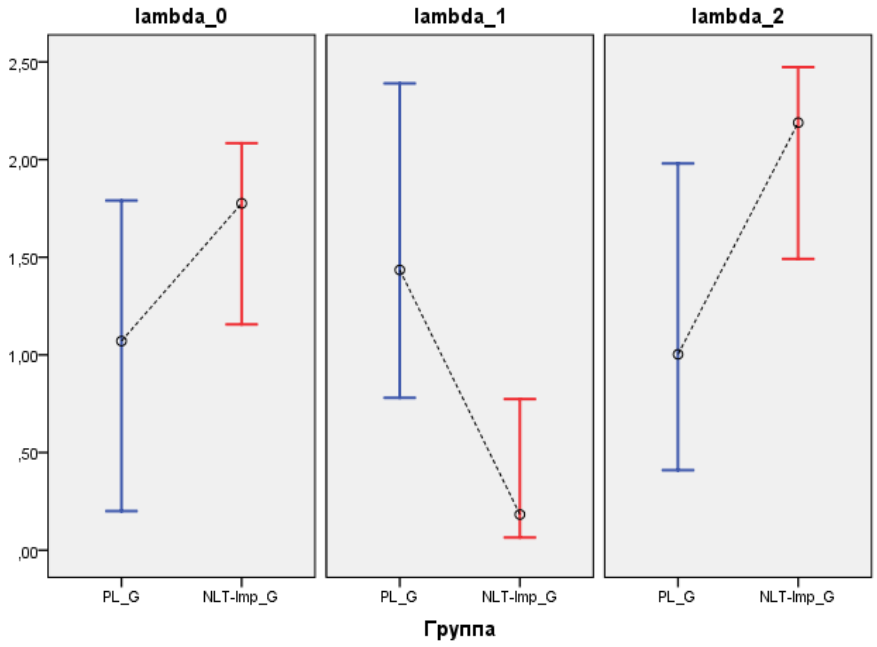

Рис. 2.5. Доверительные интервалы параметров экстенсивности АЛТ

распад тканей печени усиливается у женщин, которые были подвержены типу терапии налетрексон-имплант и у мужчин с типом терапии плацебо.

Таблица 2.16. Доверительные интервалы параметров экстенсивности АЛТ

Группа	λ_0	λ_1	λ_2
PL_G	(0,2; 1,79)	(0,78; 2,39)	(0,41; 1,98)
$NLT - Imp\_G$	(1,009; 1,996)	(0,106; 0,803)	(1,394; 2,497)

Глава 3

Итерционно-частичный метод дискриминантного анализа для неполных данных

3.1. Постановка задачи

Основной целью дискриминантного анализа является нахождение такой линейной комбинации переменных, которая бы оптимально разделила рассматриваемые группы выборки данных. Линейная функция $d_{km} = b_0 + b_1 x_{1km} + \ldots + b_p x_{pkm}, m = 1, \ldots, n, k = 1, \ldots, g$ называется канонической дискриминантной функцией с неизвестными коэффициентами b_j. Здесь d_{km} - значение дискриминантной функции для $m-$ объекта в группе k; а x_{ikm} - значение дискриминантной переменной для $m-$ объекта в группе k.

Классификация индивидов будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп. Понятно, что большая внутригрупповая вариация нежелательна, так как в этом случае любое заданное расстояние между двумя средними тем менее значимо в статистическом смысле, чем больше вариация распределений, соответствующих этим средним. Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции d, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой $\lambda = B(d)/W(d)$ где B – межгрупповая и W – внутригрупповая матрицы рассеяния наблюдаемых элементов от средних. Иногда в формуле вместо W используют матрицу рассеяния T объединенных данных.

Почти любые данные, на которых проводились исследования, являются неполными. Неполнота данных, определяется обычно объективными факторами и является одной из проблем в решении задачи построения наилучшего предсказания зависимой переменной Y по совокупности независимых переменных $X_1, \ldots, X_n$. В большинстве случаев удаление неполных наблюдений приводит к уменьшению размера выборки до непригодного к исследованию либо приводит к потере информации. В связи с этим, возникла идея на базе дискриминантного анализа и построения дискриминантной функции по всем переменным, рассмотреть совокупность наиболее значимых частичных предсказа-

ний (дискриминантных функций), построенных по подмножествам независимых переменных. Подобное частичное предсказание уже было выведено на базе множественной регрессии и, опираясь на полученные результаты, можно сделать предположение, что в случае построения частичных дискриминантных функций и использования их взвешенного усреднения, можно получить оптимальное приближение к полному предсказанию. Тем самым научиться классифицировать каждого из индивидов исследуемой базы данных.

Таким образом, пусть $X = (X_1, \ldots, X_n)^T$ - вектор независимых случайных величин. Тогда задача состоит в том, чтобы предсказать переменную Y по независимым переменным $X_1, \ldots, X_n$ в случае неполных данных или другими словами - предсказать переменную Y с помощью n линейно независимых частных дискриминантных функций.

3.2. Алгоритм анализа

Строим на полной выборке данных дискриминантную функцию успешно разделяющую индивиды на 2 группы результата завершения программы лечения по всем показателям. Далее разделим выборку на группы таким образом, чтобы в группах, по возможности, отсутствовали одни и те же показатели. При этом, стоит оставить возможность, управлять размером группы в зависимости от полученных коэффициентов взвешенного усреднения и точности классификации по частным дискриминантным функциям. Далее внутри каждой группы строим частные дискриминантные функции. После чего возвращаемся к полной выборке и по полученным частным дискриминантным функциям строим полную.

3.3. Описание эксперимента

Исследование проведено на базе отдела наркологии Национального Медицинского Исследовательского Центра Психиатрии и Неврологии (НМИЦ ПН) им.В.М.Бехтерева в период 2012˘2017гг. В рамках двойного слепого плацебо-контролируемого исследования 150 пациентов с синдромом алкогольной зависимости (МКБ-10) были случайным образом распределены (рандомизированы) на 3 группы:

1. пациенты первой группы (50 человек) получали дисульфирам
2. пациенты основной группы (50 человек) получали цианамид

3. пациенты группы сравнения (контрольная группа) (50 человек) получали плацебо идентичного вида.

Препарат для исследования назначался на 3 месяца (12 недель), в течение которых испытуемые должны были посещать исследовательский центр еженедельно для контроля ремиссии, соблюдения режима приема лекарств (при наличии флуоресцентного маркера рибофлавина в моче), а также для психометрические оценки. Кроме того, у пациентов были взяты образцы крови. Различия в продолжительности удержания в программе лечения (т.е. в ремиссии) у пациентов с разным типом терапии были выявлены посредством анализа выживаемости Каплана-Мейера. Значимость различий в кривых выживаемости оценивали с использованием критерия Гехана-Вилкоксона. В качестве шкалы классификации - время до выхода из программы (время до выхода из программы по любой причине, рецидив, нарушение условий участия).

Целью исследования является разработка математического метода выявления совокупности наиболее значимых частичных предстказаний (дискриминантных функций) и выражения полного предсказания через частичные.

Методы решения поставленной задачи: построение полной дискриминантной функции и совокупности наиболее значимых частичных дискриминантных функций, проверка на прикладных данных гипотезы о возможности через взвешенное усреднение частичных дискриминантных функций получить полную, а так же вывод алгебраической формулы наилучшего линейного предсказания для дискриминантного анализа с явно указанным видом корреляционной матрицы.

В качестве признаков были 44 показателя:

1. характеризующие давность и тяжесть заболевания алкоголизмом у индивида
2. демографические
3. характеризующие семейное состояние и наличие заболевания у родственников
4. общие показатели алкоголизации за период лечения

3.4. Описание результатов

Для начала исследования из всего количества признаков было выбрано произвольно 14 При построении полной дискриминантной функции на заявленных данных полу-

чаем значение канонической корреляции равное $0,924$, а так же подтвердить значимое отличие в обеих группах по результату завершения программы средних значений дискриминатной функции ($p - value = 0,0001$). А большое собственное значение $5,835$ указывает на хорошо подобранную дискриминатную функцию.

Средние значения дискриминантной функции в обих группах:

Таблица 3.1. Средние значения дискриминантной функции в обих группах

Группа	Функция
Програма завершена	$2,820$
Выбывание из программы	$-2,042$

Средее значение внедиаганальных элементов корреляционной матрицы равно $-0,000059$. И по матрице классификаций получилось, что 92% индивидов классифицированы корректно. А коэффициенты дискриминантной функции получились следующие:

Таблица 3.2. Коэффициенты полной дискриминантной функции

X_1	X_2	X_3	X_4	X_5	X_6	X_7
$0,167$	$0,009$	$-0,083$	$0,045$	$-0,005$	$0,092$	$-0,080$
X_8	X_9	X_10	X_11	X_12	X_13	Константа
$0,189$	$0,001$	$-0,005$	$-0,001$	$-0,192$	$0,482$	$-3,251$

Далее разделим выборку таким образом, чтобы в группах, отсутствовали одни и те же показатели. Далее путем перебора частных дискриминантных функций для поиска наилучшего предсказания полной дискриминантной функции путем, например, перебора по возрастанию - сначала по 1 признаку, потом по комбинации двух и так далее, пока не получим требуемое количество значимых частных дискриминантных функций наилучшим образом приближающих полную дискриминантную функцию. Сравнение предсказаний можно выполнять по метрике различия - например, квадрат расстояния Махаланобиса или лямбда Уилкса. По результатам построения частных дискриминантных функций имеем следующие результаты:

Таблица 3.3. Коэффициенты первой частной дискриминантной функции

X_2	X_4	X_6	X_8	X_10	X_12	Константа
$-0,019$	$-0,016$	$-0,030$	$-0,465$	$0,003$	$0,503$	$1,009$

Средее значение внедиаганальных элементов корреляционной матрицы равно $0,189$, канонической корреляцией $0,759$. И по матрице классификаций получилось, что 86% индивидов классифицированы корректно.

Таблица 3.4. Коэффициенты второй частной дискриминантной функции

X_1	X_3	X_5	X_7	X_9	X_11	X_13	Константа
$0,258$	$-0,075$	$-0,014$	$-0,002$	$0,001$	$-0,005$	$0,539$	$-3,520$

Средее значение внедиаганальных элементов корреляционной матрицы равно $-0,0132$, канонической корреляцией $0,625$. И по матрице классификаций получилось, что 82% индивидов классифицированы корректно.

Если вернуть на из первоначальные места компоненты построенных частных дискриминантных функций, то можно построить полную дискриминантную функцию по полученным частным 3.5. Процентное соотношение отличия коэффициентов двух матриц представлено на рисунке 3.1.

Таблица 3.5. Коэффициенты полной дискриминантной функции

X_1	X_2	X_3	X_4	X_5	X_6	X_7
$0,258$	$0,019$	$-0,075$	$0,016$	$-0,014$	$0,030$	$-0,002$
X_8	X_9	X_10	X_11	X_12	X_13	Константа
$0,465$	$0,001$	$-0,003$	$-0,005$	$-0,503$	$0,539$	$-2,511$

Сравним теперь результаты классификации по полной дискриминантной функции и по полученным частным. Если в случае полной дискриминантной функции процент предсказания был найден как 92%, то в случае перебора частных дискриминантных функций данный показатель снизился до 84%.

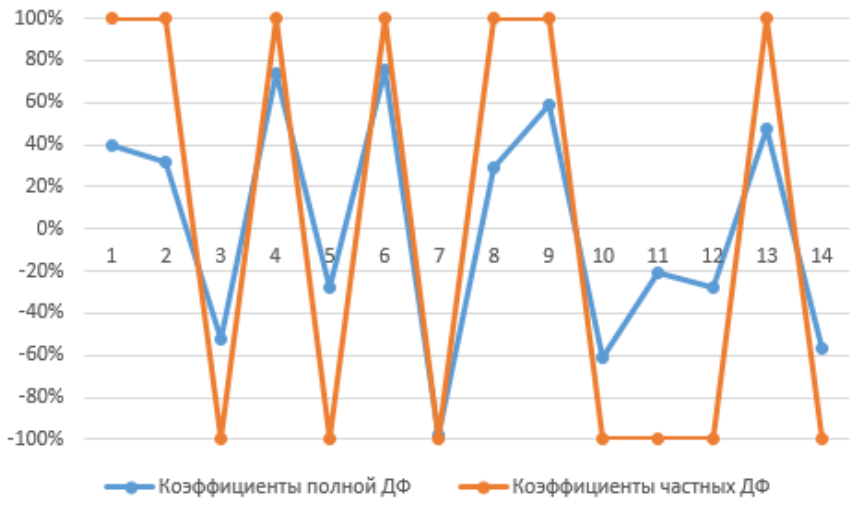

Рис. 3.1. Нормированный график процентного соотношения коэффициентов построенных дискриминантных функций

Стоит заметить, что чем больше переменных используется в полной дискриминантной функции, тем лучше получается результат классификации частными функциями в приближении к полной.

На графике 3.2 представлены результаты процентов классификации полной и частными дискриминантными функциями

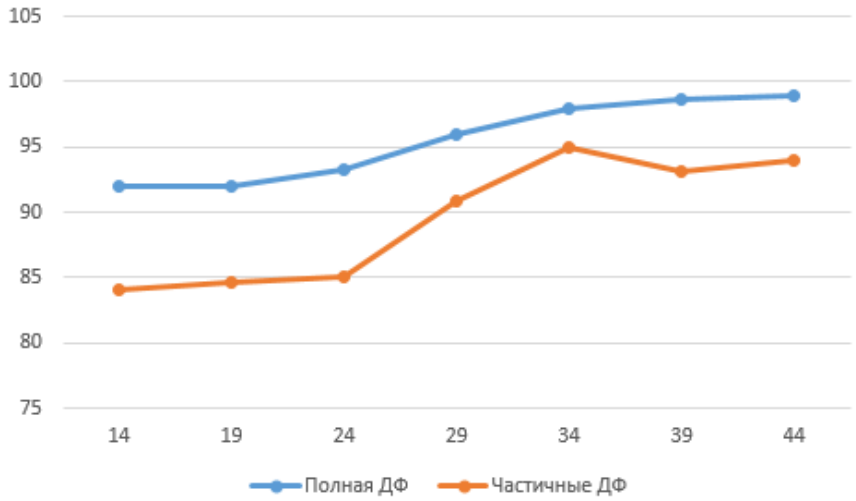

Рис. 3.2. График результатов классификации

Заключение

В данной работе были рассмотрены математические методы исследования прикладных данных.

Работу можно разделить на 3 части:

1. Симптомный, супер-симптомный метод показал, что в случае многомерного анализа данных, когда отдельные факторы незначительны, можно выявить группу риска с помощью специальной комбинации факторов. С помощью статистического пакета R был реализован алгоритм отбора сочетаний генов, влияющих на результат выполнения программы лечения, а также набор функций, который позволяет автоматизировать разработанный метод исследования. Программа применима к любым категориальным данным, что делает ее полезным приложением в генетике и дргуих сферах.

   Программа уже была использована в рамках гранта на изучения связи между генетическими факторами больных с алкогольной зависимостью и их лечения медицинским препаратом.

   На базе полученных результатов в соавторстве с научным руководителем Алексеевой Н.П. и аспиранткой 3 курса AL-JUBOORI, Fatema Saik подготовлены и отправлены в публикацию 3 статьи.

   С результатами проведенного исследования успешно состоялось выступление на 10th International Workshop on Simulation and Statistics. Тезисы были успешно опубликованы и представлены слушателем. Слушатели отметили интересный подход к исследованию данных, а так же обратили внимание на новизну предлагаемого решения.

2. Применение двумерного гамма-распределения к анализу данных. Данный подход показал, что в случае, когда отсутствую значимые различия между группами по средним, еще не означает, что различий нет вовсе. Данный метод позволяет проверить для каких подмножеств выборки тот или иной фактор является значимым и при необходимости определить направление изменения показателей с течением времени, в отличие от установки значимого влияния посредством других известных методов проверки однородности.

С помощью статистического пакета R была выполнена проверка согласия с гамма распределением, выполнена проверка однородности параметров двумерного гамма-распределения, оценка параметров гамма-распределения разными способами, исследована плотность распределения и получены доверительные интервалы для оценок параметров распределения, значимость и направленность изменения признаков в зависимости от различных факторов.

Программа была использована для поиска связи между психометрическими признаками и генетической палитрой больных с алкогольной зависимостью.

3. В связи с тем, что неполнота данных обычно обусловлена объективными факторами и осложняет решение задачи построения наилучшего линейного предсказания зависимой переменной Y по комплексу независимых переменных $X_1, \ldots, X_n$. Идея предполагаемого метода заключается в том, чтобы вместо одного выражения дискриминантной функции, построенного сразу по всем переменным (полное предсказание), рассматривать совокупность наиболее значимых частичных предсказаний, построенных по разным подмножествам независимых переменных. Отсюда возникает задача выражения полного предсказания через частичные, результаты исследования которой приведены в рамках данной работы.

Далее планируется завершить аналитическое представление полной дискриминантной функции через частные, а также представить метрику различия для сравнения частных дискриминантных функций и полной.

Список литературы

1. Алексеева Н. П. Анализ медико-биологических систем. Реципрокность, эргодичность, синонимия. — Издательство С.-Петербурского университета, 2012. — 184 с.

2. Бородин А. Н. Элементы теории вероятностей и математической статистики. — Издательство «Лань», 1999. — 224 с.

3. E.A. Gehan (1975). Statistical methods for survival time studies. In Cancer Therapy: Prognostic Factors and Criteria. M. J. Staquet, ed., pages 7–35. New York: Raven Press.

4. N.Alexeyeva, P. Gracheva, B. Martynov, I. Smirnov (2009). The finitely geometric symptom analysis in the glioma survival. In The 2nd International Conference on BioMedical Engineering and Informatics ($BMEI09$). Okt.2009., page DOI: 10.1109/$BMEI$.2009.5305560. China.

5. Mathai Arak M, Moschopoulos Panagis G. On a multivariate gamma // Journal of Multivariate Analysis. — 1991. — Vol. 39, no. 1. — P. 135–153.

6. N.P. Alexeyeva, E.P. Skurat. Symptom analysis of multidimensional categorical data with application in genetics In The 10th International Workshop on Simulation and Statistics. Sep.2019., page 89 . Salzburg, Austria.

7. N.P. Alexeyeva, F.S. Al-Juboori, E.P. Skurat. Symptom analysis of multidimensional categorical data with applications // Periodicals of Engineering and Natural Sciences. — 2020. — Vol. 8, no. 3. — P. 1517–1524.